MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies

Loes Hollestein[1,2], Serigne N Lo[3,4,5], Jo Leonardi-Bee[6], Saharon Rosset[7,8], Noam Shomron[8,9], Dominique-Laurent Couturier[10,11], Sonia Gran[12]

1 Erasmus MC University Medical Center, Dept. Dermatology, Rotterdam, The Netherlands
2 Netherlands Comprehensive Cancer Organization (IKNL), Dept. Research, Utrecht, The Netherlands
3 Melanoma Institute Australia, The University of Sydney, Sydney, NSW, Australia
4 Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia
5 Institute for Research and Medical Consultations (IRMC), Imam Abdulrahman Bin Faisal University"
6 Centre for Evidence Based Healthcare, Division of Epidemiology and Public Health, University of Nottingham, UK
7 Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel
8 Edmond J Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv, Israel
9 Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
10 Medical Research Council (MRC) Biostatstics Unit, University of Cambridge, United Kingdom
11 Cancer Research UK – Cambridge Institute, University of Cambridge, United Kingdom
12 Center for Evidence-Based Dermatology, School of Medicine, University of Nottingham, Nottingham, United Kingdom

Words: 1537

Correction for multiple comparisons in general

Research articles typically present the results of several hypothesis tests and often state 'All tests with p-values <0.05 were considered statistically significant'. Ignoring that multiple tests were performed induces false positive findings. Indeed, when multiple true null hypotheses are tested, the probability of rejecting at least one null hypothesis (referred to as the overall Type-I error rate or family-wise error rate [FWER]) increases with the number of tests. For instance, if 20 independent statistical tests are performed at the 0.05 significance level in a scenario in which all null hypotheses are true, the probability of rejecting at least one null hypothesis is almost 65%. This inflation of the Type-I error rate, known as multiple testing problem or multiplicity, constitutes a real challenge to researchers and partly explains the lack of reproducibility of scientific findings [1]. Many procedures have been developed to overcome multiplicity [2]. Due to its simplicity, the most widely used approach is the Bonferroni procedure, where Type-I error for each test equals the target overall Type-I error level (usually 0.05) divided by the number tests. This multiplicity correction leads to a FWER close to the target overall Type-I error level when all tests are independent but is known to be overly conservative when the tested hypotheses are related, leading to an unnecessary loss of power (i.e., lower probability of finding true associations). Therefore, multiplicity correction methods taking their dependence into account are generally preferred to gain power (e.g. resampling methods such as bootstrap and permutation tests) [3,4]. When the number of tests is very large, like in omics studies (e.g. genomics or transcriptomics), the control of the false discovery rate (FDR, i.e. the proportion of true null hypothesis among all rejected null hypotheses) is usually preferred to the control of the

FWER as it allows notable gains in power [5]. The choice regarding which method to use depends on the type of study and the hypotheses to be tested. This editorial aimed to briefly discuss multiplicity correction used in different contexts and state BJD multiplicity requirements for publication.

Clinical Trials

Sample sizes of clinical trials are based on a single or co-primary endpoints [6,7]. A trial with co-primary endpoints is considered negative if the result related to any of the co-primary endpoints is not significant. The use of multiple primary endpoints for a given sample size induces a loss of power but does not increase the Type-I error rate. In addition to the primary endpoint(s), a set of secondary and exploratory endpoints, for which no *a priori* sample size calculation was performed, is usually tested as well. In order to prevent false positive findings among the set of secondary endpoints, a clear distinction between the true secondary endpoints (i.e. which may support the primary endpoint and/or show additional effects after success of the primary endpoint) and the exploratory endpoints (i.e. hypothesis generating or endpoints with very low event rate) should be made [7]. Hypothesis testing for exploratory endpoints is not recommended [6], but the Type-I error rate should be controlled for secondary endpoints, typically by means of a FWER approach. If there is no effect on the primary endpoint(s), no effect on related secondary endpoints may be expected so that one may need to decide to stop statistical testing after a non-significant result (a fixed sequence or serial gatekeeping approach)[8]. Endpoints may also be grouped into families (e.g. a family of multiple effectiveness outcomes and a family of multiple quality of life scores). All endpoints within a family can be tested with a correction for multiple comparisons and one may only proceed to the next family when there is statistical success in the preceding family (a fixed sequence approach applied to families).

Omics studies

Omics studies investigate the relationship between a particular type of sample molecules and a sample attribute. Examples are Genome Wide Association Studies (GWAS) in which a large set of Single Nucleotide Polymorphisms (SNP) is tested for the association with an outcome of interest (e.g. skin cancer) or RNA-Seq experiments in which differences in gene/protein expression between conditions (e.g., treated vs not treated) are investigated.

As such studies typically involve hundreds to millions of (usually dependent) simultaneous tests, a FWER control of the Type-I error would lead to a drastic loss of power, explaining why FDR approaches, controlling for the fraction of false discoveries among the rejected hypotheses, are preferred [9]. The most often used FDR multiplicity correction is the one introduced by Benjamini and Hochberg (BH) and is valid for independent [10] or *positively* dependent test statistic [11], like, for example test statistics (positively) correlated due to measurement errors affecting all or some parameters of interest in a common way. As other dependence structures may be observed in practice, a FDR approach valid under more general dependence structures was later introduced by Benjamini and Yukutieli (BY) at the price of some loss of power [11].

Subgroup analyses

False positive findings may occur in studies where subgroup analyses are performed without multiplicity adjustment (e.g. a meta-analyses stratified by time-points of an outcome). As tests of such analyses typically involve correlated outcomes and/or comparisons repeatedly involving the

same groups, a resampling-based FWER multiplicity correction would lead to the greatest power. To maintain a high power, a limited number of subgroup analyses should be pre-specified in the protocol, where the subgroups chosen should be based on a clear hypothesis with a pre-existing biological rationale.

Regression analyses

If regression models are used for causal inference, hypotheses of the association between an exposure and outcome are tested and multiplicity should be addressed, if there is more than one outcome, using the methods mentioned above. Note that in parametric models (e.g. generalised linear models and survival models), the dependence between the tests of interest can usually be obtained under standard asymptotic normality assumptions, allowing the dependence between them (e.g. middle age vs young age and old age vs. young age) to be taken into account when performing FWER multiplicity corrections[2]. This leads to gain in power compared to Bonferroni-like multiplicity corrections.

When developing prediction models, the number of subjects (linear regression), cases (logistic regression) or events (survival models) determines the amount of statistical power and thus how many variables can be included in the model [12,13]. As a rule of thumb 10 subjects/cases/events are needed per variable. When developing a prediction model with a multiplicity of variables and a too low number of events, there is a risk of predicting random error (i.e. overfitting) and very poor performance of the prediction model in another patient sample. In those situations even more than 10 subjects/cases/events per variable may be required [14].

Correction for multiple comparisons in the BJD

Multiple comparisons can be foreseen at the design phase of the study, when multiple hypotheses are formulated. Therefore, methods to correct for multiple comparisons should be pre-specified in the protocol and/or the statistical analysis plan. The BJD requires that clinical trials and systematic reviews are pre-registered and encourages that the protocols of trials are published elsewhere and submitted as a supplementary file. We encourage authors of any type of study to consider multiple testing strategies before the start of the study and clearly report the strategy of choice in the methods.

References

1 "Reproducibility of Science: p-Values and Multiplicity" in Eighth International Purdue Symposium on Statistics, Berger JO, http://www.stat.purdue.edu/symp2012/docs/Purdue_Symposium_2012_Jim_Berger_Slides.pdf, Last accessed: ^7 June 2021

2 Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. Boca Raton, Florida, USA: Chapman and Hall/CRC. 2011.

3 Dudoit S, Popper Shaffer J, Boldrick JC. Multiple Hypothesis Testing in Microarray Experiments. *Statist. Sci.* 2003; **18**: 71-103.

4 Westfall PH, Troendle JF. Multiple testing with minimal assumptions. *Biom J* 2008; **50**: 745-55.

5 Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J Clin Epidemiol* 2014; **67**: 850-7.